LANGUAGE SYSTEMS, INC.
MUC-3 TEST RESULTS AND ANALYSIS[1]

Christine A. Montgomery
Bonnie Glover Stalls
Robert S. Belvin
Robert E. Stumberger

Language Systems, Inc.

6269 Variel Avenue, Suite F
Woodland Hills, CA 91367
lsi001%lsi-la.uucp@ism.isc.com
1-818-703-5034

RESULTS

Table 1 summarizes LSI's scores for Phase II of MUC-3. In evaluating these scores, which place LSI about two/thirds of the way down in a results-ranked list of MUC-3 participants, it should be noted that our MUC-3 system reflects a major redevelopment of the key components of the DBG message understanding system, which is currently in process. Specifically, innovative development of a parser based on government-binding principles is under way, [2] with associated revisions of the lexicon, functional parse (recovery of the predicate/argument functions representing the underlying semantic structure of the sentence), and DBG template[3] generation and frame hierarchy components (the areas indicated by the heavy lines in the system flow chart shown in Figure 1).

This innovative development is described more fully in the system summary paper. For the purposes of this site report, it is obvious that the "under construction" status of the DBG system had considerable impact upon our ability to achieve a respectable score. Had we chosen instead to go with the fairly robust previous version of the DBG system (described in [2] and [3], recently evaluated for Rome Laboratory by KSC, Inc.), our MUC-3 scores would certainly have been substantially better, because all components of the DBG system would have been fully functional (see the discussion on functionality of the DBG version currently under development below).

However, we felt strongly that the time had come to replace our chart parser with weighted rules by a more powerful and generic model that would provide a better foundation for current work, including automated translation and the integration of speech processing with the DBG system, as well as for the complex MUC-3 messages. Once the decision was made to embark upon this major re-development of the key DBG system components, it would have been unproductive to carry out the MUC-3 development in parallel on the older version of the DBG system (as well as infeasible given LSI's available resources for MUC-3).

ALLOCATION OF EFFORT

For the reasons described above, the bulk of our development effort was concentrated on the parser (approximately 60% of the total MUC-3 development effort) and the lexicon (approximately 30% of the total effort),

[1] The work reported in this paper was supported in part by the Defense Advanced Research Projects Agency, Information Systems and Technology Office, under Contract No. MDA903-89-C-0041 (Subcontract MC 1350)

[2] Unlike chart parsers, which are based on a well-understood model that has been in existence for almost 30 years, very few attempts have been made at implementing GB parsers (see [1]), and LSI's implementation approach differs from these in several important respects.

[3] It is important to note that the term "template" in the DBG system is a label for the generic message level semantic and pragmatic representational units, not an application oriented structure like the MUC-3 templates. It is the glass box output or internal representational output, as opposed to the MUC-3 templates, which are black box outputs mapped to the external representation required by a given application: currently, extraction of specific data elements for describing terrorist incidents in 9 Latin American countries.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **1991**	2. REPORT TYPE	3. DATES COVERED **00-00-1991 to 00-00-1991**

4. TITLE AND SUBTITLE **MUC-3 Test Results and Analysis**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language Systems Inc,6269 Variel Avenue,Woodland Hills,CA,91367**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		**8**	
unclassified	**unclassified**	**unclassified**			

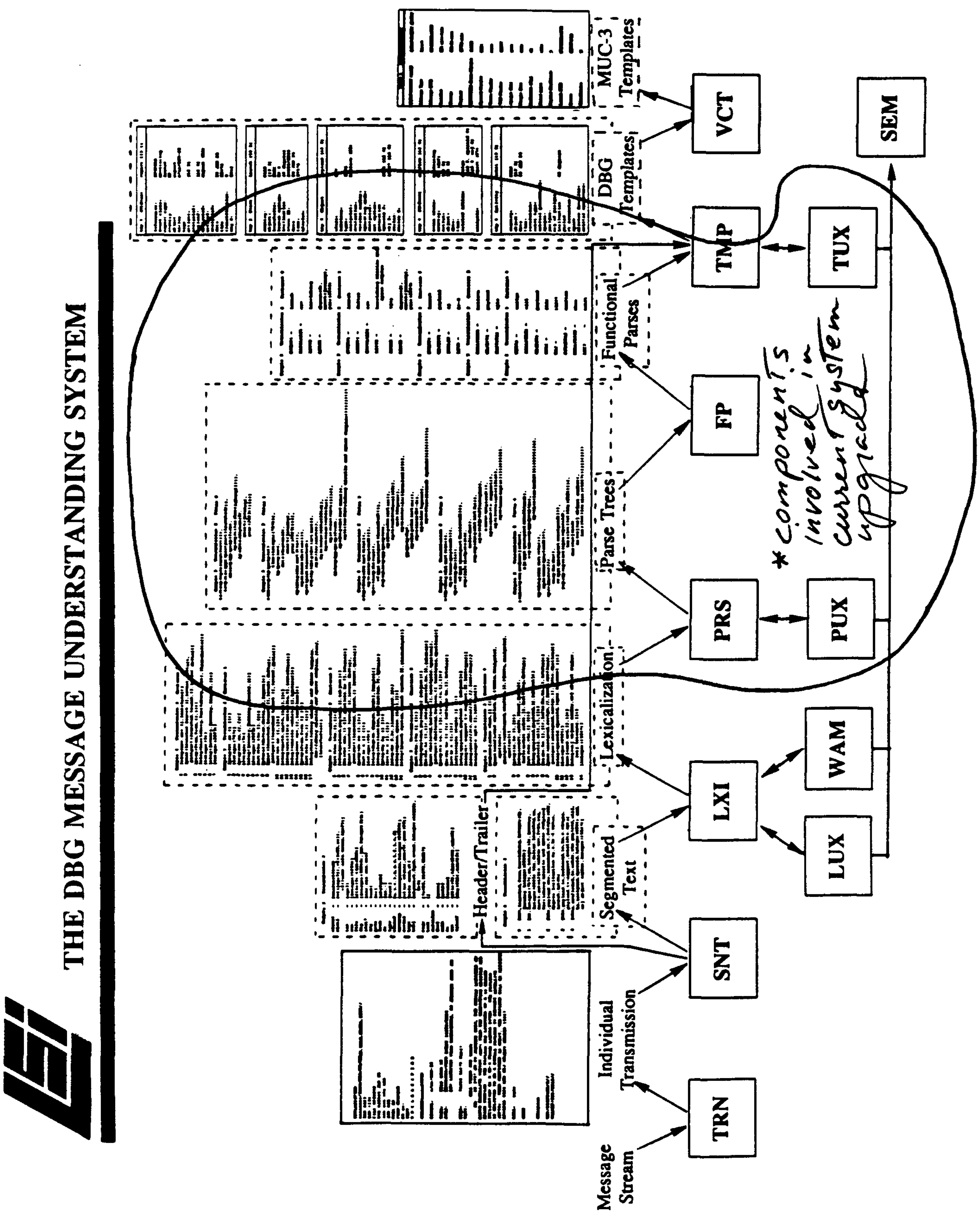

Figure 1: Functional Flow of the System

SLOT	REC	PRE	OVG	FAL
MATCHED ONLY	32	48	22	
MATCHED/MISSING	16	48	22	
ALL TEMPLATES	16	33	46	
SET FILLS ONLY	15	45	32	0

TABLE 1: SUMMARY OF SCORES

FALSE POSITIVE	FALSE NEGATIVE	
Message 30	Message 01	Message 46
Message 38	Message 08	Message 54
Message 66	Message 16	Message 56
Message 83	Message 18	Message 57
Message 84	Message 19	Message 68
Message 85	Message 24	Message 74
Message 88	Message 28	Message 78
Message 92	Message 31	Message 80
Message 96	Message 32	Message 90
	Message 36	Message 93
	Message 39	Message 95
	Message 42	

TABLE 2: INCORRECTLY CLASSIFIED MESSAGES (RELEVANT VS. IRRELEVANT)

Message 07
Message 10
Message 11
Message 16
Message 23
Message 30
Message 67
Message 77
Message 86
Message 91
Message 99

TABLE 3: UNPROCESSED MESSAGES

since our earlier system of categories and features had to be substantially revised and enhanced to provide all the subcategorization and selection information required by the GB parser. The remaining 10% of the effort was devoted to the higher level components for semantic and pragmatic interpretation at the sentence and message level (i.e., the functional parse, (DBG) template generation and frames component, and back-end application component for extracting data for filling MUC-3 templates.

Since December, approximately 10,000 new lines of commented Prolog code have been written, and are in the process of being debugged. The lexicon developed for MUC-3 comprised almost 10,000 entries including inflected forms. In comparison, the unclassified lexicon for the air activities messages recently completed for Rome Laboratory [3] included only about 1500 entries, and the most recent lexicon for the Army maneuver dialogs is estimated at about 3,000 entries. The MUC-3 frames component includes 3485 frames and 257 classes of items. However, much of this information was not exploited properly since, as mentioned above, very little of the MUC-3 development effort could be devoted to the higher level DBG components which utilize frame and slot information.

LIMITING FACTORS

In LSI's case, these included all of the items on the NOSC list of possible limiting factors, i.e., time, people, cpu cycles, and knowledge (interpreted as limitations on system use of available knowledge), as discussed below.

Time/People

There were not enough person hours available to LSI under the given resource limitations to carry out the full effort required for MUC-3 in the time allotted, as well as to perform on other in-house contracts. Although particular sentence types represented in the MUC-3 texts were no more complex than in previous projects such as the air activities work carried out for Rome Laboratory, there was a great deal more variety in the sentence types in the MUC-3 texts, and the time and resources available were substantially less than the 18 months and $225K dedicated to the Rome effort.

Cpu Cycles

The main problem here is that we were essentially developing and debugging a substantial number of lines of new code, requiring detailed tracing facilities to identify and fix bugs. In addition, a key component for MUC-3 texts is LUX (Lexical Unexpected Inputs) and its associated module WAM (Word Acquisition Module) which deal with items not present in the DBG lexicon, either because they are misspellings of existing entries, or because they are new words. LUX in particular consumes a lot of cpu cycles, but is absolutely critical for processing of texts containing many words which are unknown to the DBG lexicon.

In order to complete Test 2 in some finite time frame, it was therefore necessary to limit the input to the DBG message understanding system by utilizing some fairly draconian measures. Discourse analysis of the development corpus revealed that detecting the transitions from descriptions of one event to another in the texts was too complicated to attempt within the limited resources, so all messages labeled as relevant were arbitrarily truncated to 10 sentences. Since the intent was to exclude all but event reports, which typically describe the event(s) of interest within the 10 sentence segment, it did not appear that much information loss would result from this measure. No attempt was made to determine the number of MUC-3 templates represented in the truncated text that began at the 11th sentence of all messages in the critical event directory. (For further discussion, see LSI's section in the paper on discourse processing within the MUC-3 context included in this proceedings.)

The selection of relevant messages, which was performed using Logicon's Message Dissemination System (LMDS), [4] was thus executed to exclude all potentially windy messages such as reports of political speeches, interviews, and clandestine radio broadcasts of political propaganda, as well as military attack events and drug traffic related events not involving terrorist acts. Critical event and irrelevant message criteria were defined in terms of LMDS profiles and used to filter the test message set into the groups shown in Table 2. The LMDS

[4] LMDS does shallow text analysis based on boolean combinations of key words/phrases with proximity and other criteria, which operate as search logic specifications within particular user-defined zones of messages (e.g., mediasource zone, with typical contents such as "Radio Venceremos", "Television Peruana", etc.).

filtering was performed as a pre-processing run which took less than a minute for the total test set of 100 messages. In Table 2, false positives (in terms of the MUC-3 test key) in the critical event directory and false negatives in the "nohit" directory are indicated by Xs.

Table 3 contains numbers of the messages which were partially processed, but produced no template output because processing hung up in the parser. Per the instructions for running the test procedure, processing was restarted with the next relevant message following these parser failures.

Knowledge Availability vs. System Functionality

The DBG system was fairly well primed with knowledge, as can be seen from the size of the lexicon and the frames data bases cited previously. However, because the the GB parser was not completely functional (in fact, was still undergoing extensive debugging, as mentioned above), many attachments were not being made, resulting in a large number of partial trees which could not labeled with their thematic roles (i.e., agent, patient, etc.). The consequences of these attachment failures propagated throughout the remainder of the processing components, resulting in predicate/argument functions unlabeled, unindexed, or missing in the functional parse, so that the DBG templates were extremely sparse, as were the MUC-3 application templates. Figures 2 and 3 show partial output for Test 2 Message 100, which illustrates this point. Essentially because of the limited functionality of the GB parser at this stage of development, a great deal of the knowledge represented in the frame hierarchy and associated rules, as well as knowledge represented in the rules for filling the MUC-3 templates, was never exploited by the system.

TRAINING

It goes without saying that the development corpus was extremely useful in lexical, syntactic, semantic, and discourse analysis for system development. We also found the treebank analysis of MUC-3 messages very useful for identifying the multitude of possible variations on a single syntactic theme. Due to the partially functional status of the evolving GB parser, we were unable to fully exploit the 1300 messages in testing for this phase of MUC-3, but were limited to a few messages that we used for regression testing. The MUC-3 corpus is a very valuable archive that we intend to utilize more fully in the next few months, as our parser stabilizes and we can take advantage of the variety of texts represented in the MUC-3 collection.

MODULE MOST OVERDUE FOR REWRITING

Most of the components of the DBG system have been at least revised and extended, and in most cases, completely replaced, as part of our evolutionary design philosophy (see the system summary paper for a discussion).

As noted previously, however, one of the oldest modules in the system is the Lexical Unexpected Inputs (LUX) module, and its associated Word Acquisition Module (WAM). These modules attempt to determine whether an unexpected lexical input (i.e., one which is not present in the lexicon) is erroneous (e.g., a misspelling of an entry actually present in the lexicon), or entirely new. In the first case, LUX goes through an elaborate procedure to determine whether a spelling error exists, which is corrected if a reasonable hypothesis for an association with a word in the lexicon can be found (e.g., in test 1, the form "kidapped" was corrected to "kidnapped"). If no correction can be made, the form is determined to be new, which requires WAM to provide a temporary grammatical category assignment so that the sentence containing the new word can be parsed.

As noted previously, our lexicon included approximately 10,000 words; however, the vocabulary in the MUC-3 development corpus is estimated at 20,000 words (see Hirschman paper in this proceedings). Clearly, the LUX/WAM components were of inestimable value to us in processing the test sets; it would have been impossible to run without them. On the other hand, because of the many new words encountered in the MUC-3 texts, LUX and WAM had to be used many times on every message, and because these procedures are non-optimized at present, the amount of time devoted to autonomous LUX/WAM processing was substantial.

Another module that should be rewritten for higher efficiency is LXI, which handles lexical lookup and morphological processing; however, LUX/WAM are first on the list.

```
                     SYNTACTIC PARSE OUTPUT

   Transmission 1   Paragraph 1   Sentence 1

'Dmax1'+(1.0):
Dmax(Dbar(D([the]:det),
          Amax(Abar(A([brazilian]:adj),
                    Nmax(Nbar(N([embassy]:noun)))))).

'Pmax1'+(1.3):
Pmax(Pbar(Pbar(P([in]:prep),
               Nmax(Nbar(N([colombia]:noun)))),
          Nmax(Nbar(N([colombia]:noun))))).

'Imax1':
Imax(Ibar(I)).

'Vmax1':
Vmax(Vbar(V([has]:third_pres))).

'Vmax3'+(1.5):
Vmax(Vbar(V([confirmed]:past),
          Dmax(Dbar(D([the]:det),
                    Nmax(Nbar(N([release]:noun),
                              Genmax(Genbar(Gen([of]:of),
                                            Amax(Abar(A([red]:adj),
                                                      Nmax(Nbar(N([globo]:noun),
                                                                Nmax(Nbar(N([journalist]:noun),
                                                                          Nmax(Nbar(N([carlos]:noun_name),
                                                                                    Nmax(Nbar(N([marcelo]:noun))
)))))))))))))))).

'Dmax3'+(1.0):
Dmax(Dbar(D([who]:pronoun))).

'Vcemax1'+(1.0):
Vcemax(Vcebar(Aux([was]:aux),
              Vmax(Vbar(Vbar(V([kidnapped]:pastpart),
                             Argmax(Argbar('Arg':'*empty*'))),
                        Pmax(Pbar(P([by]:prep),
                                  Amax(Abar(A([colombian]:adj),
                                            Nmax(Nbar(N(['army of national liberation']:noun),
                                                      Nmax(Nbar(N([guerrillas]:plural))))))))))))).

******************************************************************************

                 SENTENCE-LEVEL SEMANTIC INTERPRETATION

functional-parse-1:
  'MAINPRED'('1.0') = 'INDEX'('1.1')
  'DETERMINER'('1.1') = the
  'ARG'('1.1') = brazilian
  'FOREIGN_GOVT_FACILITY'('1.1') = embassy
  'DESCRIPTION'('1.1') = 'INDEX'('1.2')
  'COUNTRY'('1.2') = colombia
  'MAINPRED'('1.2') = 'INDEX'('1.3')
  'MAINPRED'('1.3') = 'INDEX'('1.4')
  'PRED'('1.4') = have
  'MAINPRED'('1.4') = 'INDEX'('1.5')
  'PRED'('1.5') = confirm
  'DETERMINER'('1.5') = the
  'MATERIAL_ACT'('1.5') = 'release of red globo journalist carlos marcelo'
  'MAINPRED'('1.5') = 'INDEX'('1.6')
  'MAINPRED'('1.6') = who
  'MAINPRED'('1.6') = 'INDEX'('1.7')
  'EVENT'('1.7') = kidnap
  'AGENT'('1.7') = 'colombian army of national liberation'
  'ORG'('1.7') = 'army of national liberation'
  'AGENT'('1.7') = guerrillas
  'MAINPRED'('1.7') = 'INDEX'('1.8')
```

Figure 2: Partial Syntactic and Semantic Parse Output for Message 100 (TST2)

```
            MESSAGE-LEVEL SEMANTIC INTERPRETATION
                      DBG TEMPLATES

Report    muc3  [1]

date:                26 may
event:               [1.1]
                     [1.2]
                     [1.3]


Action    kidnap  [1.1]

agent:               colombian army of national liberation
agent_org:           army of national liberation


Action    kidnap  [1.2]


Action    abduct  [1.3]

agent:               the guerrillas
patient:             he

******************************************************************************

                APPLICATION OUTPUT
                  MUC TEMPLATES

0.  MESSAGE ID                          TST2-MUC3-0100
1.  TEMPLATE ID                         1
2.  DATE OF INCIDENT                    - 26 MAY
3.  TYPE OF INCIDENT                    KIDNAPPING
4.  CATEGORY OF INCIDENT                TERRORIST ACT
5.  PERPETRATOR: ID OF INDIV(S)         "COLOMBIAN ARMY OF NATIONAL LIBERATION"
6.  PERPETRATOR: ID OF ORG(S)           ARMY OF NATIONAL LIBERATION
7.  PERPETRATOR: CONFIDENCE             REPORTED AS FACT
8.  PHYSICAL TARGET: ID(S)              *
9.  PHYSICAL TARGET: TOTAL NUM          *
10. PHYSICAL TARGET: TYPE(S)            *
11. HUMAN TARGET: ID(S)                 -
12. HUMAN TARGET: TOTAL NUM             1
13. HUMAN TARGET: TYPE(S)               -
14. TARGET: FOREIGN NATION(S)           -
15. INSTRUMENT: TYPE(S)                 *
16. LOCATION OF INCIDENT                -
17. EFFECT ON PHYSICAL TARGET(S)        *
18. EFFECT ON HUMAN TARGET(S)           -

0.  MESSAGE ID                          TST2-MUC3-0100
1.  TEMPLATE ID                         2
2.  DATE OF INCIDENT                    - 26 MAY
3.  TYPE OF INCIDENT                    KIDNAPPING
4.  CATEGORY OF INCIDENT                TERRORIST ACT
5.  PERPETRATOR: ID OF INDIV(S)         "THE GUERRILLAS"
6.  PERPETRATOR: ID OF ORG(S)           -
7.  PERPETRATOR: CONFIDENCE             REPORTED AS FACT
8.  PHYSICAL TARGET: ID(S)              *
9.  PHYSICAL TARGET: TOTAL NUM          *
10. PHYSICAL TARGET: TYPE(S)            *
11. HUMAN TARGET: ID(S)                 "HE"
12. HUMAN TARGET: TOTAL NUM             1
13. HUMAN TARGET: TYPE(S)               -
14. TARGET: FOREIGN NATION(S)           -
15. INSTRUMENT: TYPE(S)                 *
16. LOCATION OF INCIDENT                -
17. EFFECT ON PHYSICAL TARGET(S)        *
18. EFFECT ON HUMAN TARGET(S)           -
```

Figure 3: DBG-Templates and MUC-3-Templates for Message 100 (TST2)

REUSABILITY

The DBG system developed for the MUC-3 application is completely reusable on other applications, with the exception of the rules for deriving the output MUC-3 templates from the DBG templates, which is the backend especially tailored for the MUC-3 application. Other than that, there are a few features such as an attribute in the frame system entitled "critical event", which would not be useful in another application, but there are very few of these (another such feature does not even come to mind at this point).

LESSONS LEARNED

Since we have performed several MUC-like tasks (i.e., data extraction) as described in the system summary, as well as evaluations, the main lesson learned was not to postpone further the acquisition of an on-line dictionary such as Longmans or the OED. In any case, had we made such an acquisition for MUC-3, time and resources would have been insufficient to integrate it with the other system components and exploit it within the MUC-3 context.

With respect to evaluation, the evaluations performed by LSI on the systems described in [2] and [3] both included competitive testing in the template-filling task against a human user or simulated user of the type of information in the given domain. This type of evaluation is perhaps more difficult in the MUC case, but, based on our experience, is extremely significant for users, because it is more believable to them than a series of finely tuned scores.

REFERENCES

[1] Berwick, R. C., Principle-Based Parsing, AI TR No. 972, June, 1987.

[2] Montgomery, C. A., Burge, J., Holmback, H., Kuhns, J. L., Stalls, B. G., Stumberger, R., Russel, R. L., The DBG Message Understanding System, in *Proceedings of the Annual AI Systems in Government Conference* (1989), IEEE Computer Society Press, 1989.

[3] Stalls, B., R. Stumberger, and C. A. Montgomery (1990). Long Range Air (LRA) Data Base Generator (DBG). RADC-TR-89-366.